Shrivastava, Jolly, M.S. An Extended Mixed Inheritance Model for Detecting Major Genes Affecting Quantitative Traits. (2005)
Directed by Dr. David L. Remington. 55pp.

The objective of this research is to extend the mixed inheritance model developed by Zeng et al (2004) to detect the segregation of two major genes using phenotypic data from a half-diallel mating design. The model can be used to select parents which are segregating for major genes, both for breeding purposes and studies of adaptive evolution. The model can be used to find parents that are heterozygous for major genes, so the cost of QTL mapping just to determine whether a QTL is present can be avoided.

A Bayesian approach using Gibbs sampling was used to develop this model. Genotypes of the parents and progeny are updated using “parent blocking” in which the genotypes of the parents and progeny are updated as a block. For this study, only additive effects of major genes were taken into account.

In general, estimates of genetic parameters were accurate. When major gene effects were large ($\geq$ 0.5 phenotypic standard deviation), the genotypes of the parents along with genetic parameters were estimated correctly. However, when major gene effects were small, transformation between major genes and polygenic effects occurred frequently and heterozygotes were sometimes incorrectly identified.

Suggestions for further modifications of the model are made, including addition of dominance, epistatic, and genotype X environment interactions and modifications to improve the mixing of the chains.

AN EXTENDED MIXED INHERITANCE MODEL FOR DETECTING MAJOR GENES AFFECTING QUANTITATIVE TRAITS

by

Jolly Shrivastava

A Thesis Submitted to
The Faculty of the Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Greensboro
2005

Approved By

_____
Committee Chair

APPROVAL PAGE

This thesis has been approved by the following committee of the faculty of the graduate school at the University of North Carolina at Greensboro.

Committee Chair _____

Committee Members _____

_____

_____

_____
Date of Acceptance by Committee

_____
Date of Final Oral Examination

TABLE OF CONTENTS

Page

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

Quantitative genetics deals with the inheritance of a character that varies continuously and cannot be classified as a type or kind. For example, in Mendelian genetics we can say that eye color is blue or brown; in other words, we can categorize it. However traits like height cannot simply be classified as tall or small since there are varying degrees of height. Thus, we have to use a different approach, and quantitative genetics deals with these types of traits. These characters are known as metric characters or quantitative traits.

Quantitative traits may be affected either by a large number of genes, each having a small effect on the trait, or by a few major genes that may produce a large variation in the phenotype of the trait under study. Genes that have a large effect on the trait are also known as oligogenes and have been shown to influence a variety of traits. The presence of oligogenes has been shown in a variety of organisms, such as Drosophila, mice, domestic animals and trees (Piper and Shrimpton 1989; Tanksley 1993; Jiang et al 1994; Kaya et al 1999). However it has been shown that the estimated size of major gene or the threshold size that can be detected varies with the power of the statistical test that has been used to detect it (Elston 1992). By contrast, polygenes have

individually minute effects on phenotype values, and a noticeable difference in the phenotype is produced only when the effects of multiple genes are combined.

1.1. Polygenes vs Major Genes: The term “polygene” was given by Mather (1941) to a gene that has a very little effect on the variation of the trait and can produce a noticeable difference only when combined with other genes.

Studies conducted on wing shape in *D. melanogaster* have shown that the trait is polygenic in nature (Weber 1990, 1992). Mutation studies done on bristle number in *D. melanogaster* have shown that the trait has a polygenic basis of inheritence (Mackay et al 1992a, 1994; Fry et al 1995).

Major genes have been shown to be present in a variety of plants and animals. In maize, resistance to Southern leaf blight disease is caused by QTL or quantitative trait loci (locations of oligogenes detected by genetic mapping). It was found that 7 QTL on six chromosomes contributed to resistance and accounted for 46% of the phenotypic variation for resistance. QTL on chromosomes 1, 2 and 3 had the largest effect (Carson, 2004). In verticillium disease of potato, resistance QTL were detected on four chromosomes and one individual QTL explained 40% of the variation of the trait (Simko et al 2004).

Information about QTL in plants has come mainly from domesticated species. Studies on cultivated taxa have shown that many characters in plants are controlled by genes of large effects (Hilu 1983; Gottlieb 1984). However some studies have also been performed on variation in natural populations. Mapping studies done using molecular markers have shown that major QTL are

responsible for the differences in inflorescence architecture between maize and teosinte (Doebley and Stec 1993), and in differences in flowering times between cultivars of *Brassica oleracea* (Camargo and Osborn 1996). It has also been shown that major QTL control the differences in the morphology of flowers between two species of monkeyflower (Bradshaw et al 1995).

Light is an important factor for plant growth, affecting germination, development and flowering. Accessions in Arabidopsis show quantitative variation for hypocotyl length under different wavelengths of light (Maloof et al 2001; Botto and Smith, 2002). When recombinant inbred lines (RIL) from Columbia and Kashmir accessions were mapped using composite interval mapping (CIM) and extreme array mapping, eight QTL were identified with five localized near photoreceptor loci. Percentage of variance explained by these QTL ranged from 7.5% to 48.2% for red light effects (Wolyn et al 2004). The results obtained from CIM were similar to that reported for a RIL population derived from Ler and Cvi ecotypes (Borevitz et al 2002).

In tomato, four loci (*fw1.1*, *fw 2.2*, *fw 3.1* and *fw 4.1*) were first identified in crosses between small wild tomatoes and large cultivated species (Grandillo et al 1999). Variation at these loci can cause a 30% change in the final fruit mass. It was also found that genetic variation at the *fw 2.2* locus on chromosome 2 alone can change the size of the fruit by up to 30% (Frary et al 2000). Developmental studies conducted on nearly isogenic lines have shown that changes in fruit size associated with *fw 2.2* are prominent in later stages of the development cycle

(Nesbitt and Tanksley, 2001; Cong et al, 2002). There are wide variety of traits affected by major genes and polygenes. The *MS10* gene in tomato flowers interacts with polygenes for the efficient production of hybrid seeds (Levin et al, 1994). One or more major genes can affect seed traits, which are expressed in embryo, endosperm or cytoplasm. This scenario was found in maize where effects of the *opaque-2* gene were modified by the genetic background (i.e. polygenes) and the negative effects of *opaque-2* gene (e.g. low grain weight, susceptibility to insects and diseases) can be removed by selection of favorable alleles (Vasal et al 1980).

1.2. QTL Mapping and Analysis: The principle behind QTL mapping is that if a QTL is linked to a particular marker then we will observe a difference in the mean values of the trait among the organisms that differ in the genotypes at that marker locus (Sax 1923). DNA markers have played a very important role in the QTL mapping studies and have been used extensively in crop plants such as rice (Huang et al 1996; Lin et al 1996; Yu et al 2002). In human populations, the sib-pair approach is used to map QTL (Lange 1986; Haseman and Elston 1972; Cardon et al, 1994).

QTL mapping in can be performed using single marker analysis, interval mapping, composite interval mapping, and multiple interval mapping. Single marker analysis was the initial method used, and attempted to find a QTL by checking for the statistical association between a mapped genetic marker and

the trait value. In a simulation study Wright and Kong (1997) used log of odds (LOD) scores to detect QTL using single marker analysis.

In interval mapping (IM) first proposed by Lander and Botstein (1989) the whole chromosome is searched and the position of a QTL is estimated between the markers. IM considers that only one QTL is present in the interval, so the results can be biased when multiple QTL's are present in the region (Lander and Botstein, 1989; Zeng, 1994).

Jansen (1993) and Zeng (1994) developed the idea of composite interval mapping (CIM) in which the mapping in a particular interval is combined with multiple regression on markers in other chromosomal regions to absorb the effect of other QTL. Multiple interval mapping (MIM) uses multiple marker intervals at the same time to find multiple putative QTL. MIM has been shown powerful enough to find QTL and the epistatic interactions between them; it can also determine the genotypic values and heritabilities of the trait. MIM uses Cockerham's model (Kao and Zeng) to specify the genetic parameters and formulas of Kao and Zeng (1997) for statistical estimation.

Several statistical methods are available to map QTL including least square (LS), maximum likelihood (ML), residual maximum likelihood (REML), and Bayesian analysis. LS analysis is computationally fast and gives estimate of the QTL position but doesn't give estimates of any other genetic parameters. The REML method (Grignola et al, 1994, 1996a, b; Grignola and Hoeschele, 1996) postulates normally distributed QTL effects and provides estimates of variance

explained by QTL. ML and Bayesian analysis enable the estimation of all genetic parameters depending on the QTL model (biallelic QTL).

Using LS analysis a single 'ghost QTL' was detected when two linked QTL were actually segregating (Haley and Knott, 1992; Martinez and Curnow, 1992). Grignola and Hoeshele using REML reported the same phenomenon. The problem of detecting "ghost QTL" was removed in composite interval mapping proposed by Jansen (1993) and Zeng (1994).

1.3. Disadvantages of QTL Mapping: When using QTL mapping examines only one cross at a time. In addition to this, QTL mapping is a highly time-consuming and expensive process since one needs to score markers for the QTL in the entire progeny set and develop the linkage map if one is not available. Also, in the end it is possible that the cross selected has no major gene segregating.

Considering the above disadvantages of QTL mapping, if we can develop a method that can find the crosses in which major genes are actually segregating using the quantitative data alone, the probability of detecting QTL will increase significantly and the cost of mapping QTLs to determine their presence can be decreased. This method could also be used after mapping QTL in a particular cross to evaluate how general the results are likely to be; i.e. to examine whether QTL of the same effect occur in a different cross, and

potentially determine whether or not a different major gene is responsible for the variation of the same trait in a different cross.

1.4. Tests to detect the presence of major genes: The simplest way to detect the presence of a major gene in a population is to cross two different inbred lines to produce a progeny population and then to use statistical approaches to infer the occurrence of major genes. This has been used to find major genes in farm animals using phenotypic observations in a nested mating design (Mayo 1989; Hill and Knott 1990; Le Roy and Elson 1992; Uimari et al 1996). In plant breeding designs diallel mating designs are commonly used to evaluate the breeding values of parents (Hallauer and Miranda, 1981; Zobel and Talbert, 1984; Zeng et al, 2000) and trait distributions from diallel data can be tested for the effects of major genes. Various methods are available to detect the presence of major genes in a population: departures from normality tests to indicate the presence of major genes; methods that are based on the resemblance between parents and offspring; and complex segregation analysis.

1.4.1. Departure from the normality tests: If a major gene with a large effect is segregating, there will be a detectable departure from normality and heterogeneity of the intra-family variances (Le Roy and Elsen, 1992; Cemal, 1996; Falconer and Mackay, 1996). The presence of major genes can be detected using graphical tools such as histograms,

which may show a bimodal or multimodal distribution if one or more major genes are present. Also, we can use tests based on skewness and kurtosis to find major gene segregation.

1.4.2. Methods that are based on the resemblance between parents and offspring:

Bartlett test: The Bartlett test (Snedecor and Cochran 1983) is used to test whether K samples have equal variances or not; it is also sometimes called the test for homogeneity of variances. The test statistic involves a comparison of separate within-group sum of squares to the pooled within-group sum of squares.

$$B = (n-k)\ln s_p^2 - \sum_{j=1}^{k}(n_j - 1)\ln s_j^2$$

where $S_p^2 = \frac{\text{Sum of squares}}{n\text{-}k}$

The Bartlett test is sensitive to normality, so if the sample has a non-normal distribution, the significance can be due to non-normality of the data rather heterogeneity of variances (Le Roy and Elson 1992; Uimari et al 1996). In a study on diallel progeny populations of loblolly pine it was observed that when skewness was larger than 0.25, the Bartlett test showed a 50% increase of false rejection rate of the null hypothesis which reached 92% when skewness was 0.41 (Zeng and Li, 2003).

Log ANOVA test: The log ANOVA test is another test for homogeneity of variances (Scheffe Box test, Sokal and Rohlf 1981). In this test the observations within a group are divided randomly among sub groups and the log of variance for each subgroup is calculated. The test uses an F-statistic, which compares the among-group mean squares to the within-group mean squares of the logs of subgroup variances.

$$F = \frac{SS_{among}/(k-1)}{SS_{within}/\sum_{j=1}^{k} m_j - 1}$$

where m_j is the number of subgroups within a group j.

The log ANOVA F-statistic is tested against a critical value of F with the degrees of freedom for the numerator and the denominator as k-1 and m_j -1, respectively. If the differences in the variances among families are greater than what was expected based on the basis of average variance within families, then the F-test for the family effect is significant (Sokal and Rohlf 1981).

The log ANOVA test is said to be less sensitive to departures from normality than the Bartlett test (Sokal and Rohlf, 1981) and has been shown to be the best test to identify candidate populations for a half diallel mating design (Zeng and Li 2003).

Fain test or sibship variance test: The Fain test was proposed by Fain (1978) and uses the means and variances of sibships. If the parental trait values

are at the extreme ends of the population distribution then they are more likely to be homozygous, while parents with intermediate values are more likely to be heterozygous and their families will have intermediate means and large variances. This indicates the presence of at least one major gene segregating.

In a diallel progeny population of loblolly pine it was shown that the Fain test was least sensitive to skewness of data (Zeng and Li, 2003). However, in an animal progeny population with a nested mating design, the Fain test showed power similar to the Bartlett test (Le Roy and Elson, 1992) and was more sensitive to skewness (Uimari et al 1996). McCluer and Kammerer (1984) found that the Fain and Bartlett tests have an advantage in that they will not give a false indication of a major gene when there is none.

The Bartlett test and log ANOVA test had a good power for detecting major genes only when $2a \geq 1.5$. All these tests could detect the presence of major genes but they were unable to identify the parents in which major genes were segregating or which parents carry favorable alleles.

1.4.3. Complex segregation Analysis: The classical model of segregation analysis was proposed by Elston and Stewart (1971). Since then many other models have been proposed that use random variables for segregation analysis (Hopper 1989; Li and Thompson 1997; Zhang et al 2000). CSA was developed for human pedigrees and has been said to be the most powerful tool for detecting major genes (Morton and Maclean, 1974; Hill and

Knott, 1990). CSA evaluates the transmission of the trait in the pedigree and calculates the genotypic probabilities of each individual in the population. It also evaluates the inheritance patterns and the phenotypic and environmental interactions to find the genetic mechanisms underlying a specific trait. It uses maximum likelihood analysis to find the estimates of the parameters that give the highest likelihood of the observed data.

CSA has been used in the identification of genes that affect a variety of traits like body weight **(**Skaric-Juric et al, 2003) and bone mineral density (Nguyen et al, 2003). CSA has been extensively used in the identification of genes that are involved in the susceptibility to diseases like cancer. In the case of breast cancer it was found that the frequency of the breast cancer susceptibility allele is 0.0006 and the carriers have 82% higher chances of getting breast cancer as compared to non-carriers, which have only an 8% chance of getting cancer in their lifetime.

A Bayesian approach has been commonly used in complex segregation analysis and has been demonstrated by Hoeshchele and VanRander 1993A, 1993b; Uimari et al, 1996). In plants, a Bayesian approach was developed by Satagopan et al. (1996), who used a Bayes factor approach to determine the most probable number of QTLs present.

The main limitation of CSA is that it requires highly specific data in large amounts to make correct estimates of the parameters. Also complex segregation analysis assumes that there is normality in the underlying distribution; if this

assumption is violated, false detection of major loci can occur (MacLean et al 1975; Mortan et al 1984).

1.4.4. Bayesian approach: Bayesian analysis estimates the underlying distribution of the parameters based on the observed distribution. In the Bayesian approach, we start with a prior distribution, which may come either from the assessment of the relative likelihoods of parameters or some other non-Bayesian observations. We then collect the data to get the observed distribution and get the likelihood of the observed distribution as a function of parameter values. Finally the likelihood of the parameters is multiplied by the prior distribution of the parameter to estimate the posterior distribution of the parameters.

The Bayesian approach is better than the ML approach in that ML can be biased when the sample sizes are small (Weir, 1996; Richter, S.J, personal communication). Also in our case we need to optimize the maximum likelihood estimate, which in this case is not possible even by using iterative algorithms e.g. EM.

The Bayesian approach has been successfully implemented in a mixed inheritance model (MIM) to find major genes in loblolly pine using a diallel-mating design with 6 parents (Zeng et al, 2003). Using simulated data, Zeng et al (2003) were able to identify the parents that were carrying major genes, including the genotypes of the parents and estimates of parameters like major additive genotypic effects. However the statistical model for MIM considers the presence

of only one major gene, which is unrealistic since there may be two or more major genes affecting the trait. Studies conducted on Drosophila have shown that different parents can be heterozygous for different QTL (Mackay, 2001). A study conducted on inflorescence traits in *Arabidopsis thaliana* showed that two different crosses had different numbers of QTL for various traits, with some QTL detected in only one of the crosses (Ungerer et al, 2003).

1.4.5. Objectives of study: This study is having two goals, first to extend the mixed-inheritance-model to allow more realistically for the segregation of two major genes and secondly test the performance of model. The objectives were as follows:

1. To modify the MGene program of Zeng et al (2003) so that it may be able to detect the segregation of two major genes.
2. To create simulated data sets with different numbers of major genes and different effect sizes on which to test the two major gene model.
3. To evaluate the ability of the modified MGene program to correctly detect major gene effects under a variety of conditions, using the simulated data.

CHAPTER II

MATERIALS AND METHODS

2.1. The modified mixed inheritance model program to detect two loci: The statistical model to specify the extended mixed inheritance model (EMIM) is given in equation 2.1.1. The explanation of the notation is given in table 2.1.1.

In the present model only the additive effects of major genes are taken into account. The dominance effects between the alleles and the epistatic interactions between the genes are assumed to be absent but can be included in future studies.

$$
\begin{aligned}
& p(\mu, m_1, m_2, u, W_1, W_2, w_{p1}, w_{p2}, f_1, f_2, \sigma_g^2, \sigma_s^2, \sigma_e^2 \mid Y) \\
\propto \quad & p(Y \mid \mu, m_1, u, W_1, \sigma_e^2) p(\sigma_e^2) p(g \mid \sigma_g^2) p(s \mid \sigma_s^2) p(\sigma_g^2) p(\sigma_s^2) p(\mu) p(a_1) p(a_2) \\
& p(W_1 \mid w_{p1}) p(w_{p1} \mid f_1) p(f_1) p(W_2 \mid w_{p2}) p(w_{p2} \mid f_2) p(f_2) \qquad (2.1.1) \\
\propto \quad & (\sigma_e^2)^{-n/2} \exp\{-\frac{1}{2\sigma_e^2}(y - X\mu - Zu - W_1 L_1 m_1 - W_2 L_2 m_2)'(y - X\mu - Zu - W_1 L_1 m_1 - W_2 L_2 m_2)\} \\
& (\sigma_e^2)^{-(\gamma_3+1)} \exp\{-\frac{\nu_3}{\sigma_e^2}\}(\sigma_g^2)^{-\frac{ng}{2}} \\
\times \quad & \exp\{-\frac{1}{2\sigma_g^2}\sum_{i=1}^{n_g} g_i^{\,2}\}(\sigma_g^2)^{-(\gamma_1+1)} \exp\{-\frac{\nu_1}{\sigma_g^2}\}(\sigma_s^2)^{-\frac{n_s}{2}} \exp\{-\frac{1}{2\sigma_s^2}\sum_{j=1}^{n_s} s_j^2\}(\sigma_s^2)^{-(\gamma_2+1)} \exp\{-\frac{\nu_2}{\sigma_s^2}\} \\
\times \quad & \exp\{-\frac{1}{2K_1^2}\mu^2\} \exp\{-\frac{1}{2K_2^2} a_1^{\,2}\} \exp\{-\frac{1}{2K_3^2} a_2^{\,2}\} \\
& \prod_{k=1}^{n} p(w_{k1} = w_{g1} \mid w_{p1a(k)}, w_{p1a(k)}\} \times \prod_{i=1}^{ng} p(w_{p1a} = w_{g1} \mid f_1) f_1^{\alpha_f - 1}(1 - f_1)^{\beta_f - 1} \\
& \prod_{k=1}^{n} p \prod_{i=1}^{ng} p(w_{p1b} = w_{g2} \mid f_2) f_2^{\alpha_f - 1}(1 - f_2)^{\beta_f - 1}
\end{aligned}
$$

Table 2.1.1 Symbols used and their definitions

Notation	**Definition**
Y	A (n×1) vector of n progeny observations
μ	The overall mean equal to μ. It can be extended to a (c × 1) vector of fixed non-genetic effects.
X	A (n×1) vector with value 1 for the overall mean of all progenies.
u	A (q×1) vector of q random polygenic effects, u'=(g', s') including n_p GCA's (g) and n_s SCA's(s).
g	g'= {g_i ,i=1,……., n_p} n_p GCAs are assumed to be mutually independent normal distributions, i.e. g| σ_g^2 ~ N (0, σ_g^2 I)
σ_g^2	GCA polygenic variance due to additive polygenic effects
s	s'= {s_i, i =1,……., n_p} n_s SCAs are assumed to be mutually independent normal distributions, i.e. s | σ_s^2 ~ N(0, σ_s^2 I).
σ_s^2	SCA polygenic variance due to dominance polygenic effects
Z	A (n×q) incidence matrix of GCA and SCA for all progenies
m_1	A (1×1) vector of major gene effect at first locus. m2'= {a1}

m_2 A (1×1) vector of major gene effect at second locus. m2'= {a2}

a_1 Additive major genotypic effect at first locus.

a_2 Additive major genotypic effect at second locus.

L_1 L1 = $\begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$, a (3×1) indicator matrix of the major gene effects for major genotypes at first locus.

L_2 L2= $\begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$, a (3×1) indicator matrix of the major gene effects for major genotypes at second locus.

W_1 An unknown (n×3) random incidence matrix of the major genotypes at first locus for n progenies.

W_2 An unknown (n×3) random incidence matrix of the major genotypes at second locus for n progenies.

W_{t1} A (1×3) row vector to form rows of W1 w1 = (1,0,0), w2 = (0,1,0) and w3 = (0,0,1).T taking values 1,2 and 3 respectively to represent major genotypes A1A1, A1A2 and A2A2.

W_{t2} A (1×3) row vector to form rows of W2 w1 = (1,0,0), w2 = (0,1,0) and w3 = (0,0,1).T taking values 1,2 and 3 respectively to represent major genotypes B1B1, B1B2 and B2B2.

e A (n×1) vector of id errors. e is assumed to be N(0, σ_e^2 I)

σ_e^2 Residual variance.

N Multivariate normal distribution

For mean and major gene effects normal priors were used:

$\mu = N(0,k_1^2)$, a1 = $N(0,k_2^2)$, a2 = $N(0,k_3^2)$

k_i^2 , for i= 1, 2 and 3 are the hyper parameters of the prior distribution.

Variance components $\sigma_g^2, \sigma_s^2 \quad and \quad \sigma_e^2$ arise from the inverted gamma distribution as follows:

σ_g^2 = $IG(\gamma_1, \nu_1)$, σ_s^2 = $IG(\gamma_2, \nu_2)$ and σ_e^2 = $IG(\gamma_3, \nu_3)$,

where $\gamma_i = 2$ and $\nu_i = (\gamma_i - 1) * \sigma_i^2$ and i = 1, 2 and 3, and

$\hat{\sigma}_1 = \hat{\sigma}_g$, $\hat{\sigma}_2 = \hat{\sigma}_s$ and $\hat{\sigma}_3 = \hat{\sigma}_e$. Conjugate beta priors were used for allele frequency at both the loci: p(f1) ~ $\beta(\alpha_{f1}, \beta_{f1})$ and p(f2) ~ $\beta(\alpha_{f2}, \beta_{f2})$.

α_{fi} and β_{fi} are equal to 1 to express prior ignorance, resulting in a uniform distribution.

Gibbs sampling was used for the estimates of the parameters. In Gibbs sampling, parameters are sampled from their posterior distributions, given the observed data and conditional on sampled values for all other parameters. The full conditional distributions are derived for each parameter, and they form the transition probabilities of the Markov chain. The parameters other than the parameter to be estimated are fixed to estimate the parameter of interest, and then the realized value of this parameter is substituted into the full conditional distributions of other parameters. It has been shown that under general conditions, the Gibbs sampling algorithm converges to the target density as the number of iterations becomes large.A parent blocking technique was used to update the genotypes of the parents and progenies. Parent blocking works by updating the genotype of the parent along with its half-sib progeny, and the genotypes of the progeny are updated twice since each progeny has two parents.(Zeng et al, 2003.)

2.1.1 Full conditional distributions of parameters: The full conditional distribution of the parameters is obtained by selecting the terms from equation 2.1.1 that contain a particular parameter. The realized value of this parameter is substituted in the full conditional distribution of other parameters whose distributions include that parameter.

The polygenic effects can be partitioned into general combining ability of each parent (GCA), specific combining ability of each parental combination (SCA), the GCA by environment interaction, and the SCA by environment interaction. For the present model only the GCA and the SCA effects were considered.

Genotypes of parents and progeny: The genotypes of the parents and the progeny were updated using "parent blocking" wherein the parental genotype is updated followed by the progeny genotype. This is done separately for the first locus and then for the second locus. The joint conditional distribution of the parent is given by:

$$p(w_{pi}, w_{i(1)}, \ldots\ldots\ldots, w_{i(ni)} \mid W_{-i(k)}, w_{-pi}, \theta, f1, f2, \sigma_g^2, \sigma_s^2, \sigma_e^2, Y)$$

where n_I is the number of parent *i*. The three possible genotypes of the progeny were summed after weighing the relative probability of each genotype to get the genotypic distribution of the parent *i*.The full conditional distribution of the parent is given by:

$$p(w_{pi} = w_T | W_{-i(k)}, w_{-pi}, \theta, f1, f2, \sigma_g^2, \sigma_s^2, \sigma_e^2, Y) \propto$$

$$p(w_{pi} = w_T | f1, f2) * \prod_{k \in i(k)} \sum_{b=1}^{3} p(w_k = w_b | w_{p1} = w_T, w_{p2}) p(y_k | w_k = w_b) \qquad (2.1.2)$$

where $y_k = y - X_k\beta - Z_k u - a2$ is the adjusted record, y is the phenotypic value of the progeny, β is the fixed non-genetic effects, Zu is the polygenic effects and $a2$ is the effect of the major gene at the other locus. The weight is given by:

$$p(\tilde{y}_k | w_k = w_T) \propto \exp\{-\frac{1}{2\sigma_e^2}(\tilde{y}_k - w_k Lm)^2\}$$

The probabilities must be normalized to ensure that the probabilities for all the three possible genotypes add up to 1; i.e.

$$\sum_{t=1}^{3} p(w_{pi} = w_T) = 1$$

The marginalized full conditional distributions of the offspring are given by:

$$p(w_k = w_T | w_{-k}, w_{-p}, \theta, f1, f2, \sigma_g^2, \sigma_s^2, \sigma_e^2, Y) \propto$$

$$p(w_k = w_T | w_{p1(k)}, w_{p2(k)}) * p(\tilde{y}_k | w_k = w_T) \qquad (2.1.3)$$

<u>*Allele frequency:*</u> Bernoulli random variables were used for the priors for the allele frequencies, with *f1* representing the prior probability of sampling an *A1* allele and (1-*f1*) the prior probability of an *A2* allele at the first locus. Similarly, *f2* is the prior probability of sampling a *B1* allele and (1-*f2*) the prior probability of *B2* allele. The posterior distributions of frequencies are sampled from a beta distribution:

$$p(f1 \mid W_1, w_{p1}, \theta, \sigma_g^2, \sigma_s^2, \sigma_e^2, Y) \propto f1^{(\alpha_{f1}+n_1-1)}(1-f1)^{(\beta_{f1}+n_2-1)} \quad (2.1.4)$$
$$p(f2 \mid W_2, w_{p2}, \theta, \sigma_g^2, \sigma_s^2, \sigma_e^2, Y) \propto f2^{(\alpha_{f2}+n_1-1)}(1-f2)^{(\beta_{f2}+n_2-1)}$$

By taking, $\alpha_{f1}=\beta_{f1}=\alpha_{f2}=\beta_{f2}=1$ we have taken the prior distribution of the frequency to be a uniform distribution.

Location parameters: The EMIM can also be written in the following way:

$$Y = H\theta + e$$

where H is the incidence matrix for the parameters and θ are the location parameters which include the mean, effects of major genes at the two loci, the 6 GCA values and the 15 SCA values. Thus, in all there are 24 location parameters. For θj, $j = 1,\ldots..,p$ the full conditional distribution of θj can be specified as:

$$\theta_j, \theta_{-j}, W_1, W_2, f_1, f_2, \sigma_g^2, \sigma_s^2, \sigma_e^2, Y) \sim N\left(\tilde{\theta}_j, \frac{\sigma_e^2}{\sum_{i=1}^{n} H_{kj}^2 + \frac{\sigma_e^2}{\sigma_j^2}}\right)$$

$$where \quad \tilde{\theta}_j = \frac{\sum_{k=1}^{n} H_{kj}\left(y_k - \sum_{r=1, r\neq j}^{p} H_{kr}\theta_r\right)}{\sum_{k=1}^{n} H_{kj}^2 + \frac{\sigma_e^2}{\sigma_j^2}} \quad (2.1.5)$$

H_{kj} is the k^{th} row and j^{th} column of the incidence matrix H, and, the same applies to H_{kr}. σ_e^2 is the error variance and σ_j^2 is the variance component for that particular location parameter. For example, if we are calculating the GCA of the parents, σ_j^2 is the variance for the GCA.

Variance components: The full conditional distribution of general combining ability variance can be specified as:

$$p(\sigma_g^2 \mid \beta, m_1, m_2, u, W_1, W_2, w_{p1}, w_{p2}, f1, f2, \sigma_e^2, Y)$$
$$\propto (\sigma_g^2)^{-\left(\frac{n_g}{2}+u_1+1\right)} \times \exp\left\{-\frac{1}{\sigma_g^2}\left[\frac{1}{2}\sum_{i=1}^{n_g} g_i^2 + v_1\right]\right\} \propto IG\left[\frac{n_g}{2}+u_1, \frac{1}{2}\sum_{i=1}^{n_g} g_i^2 + v_1\right] \qquad (2.1.6a)$$

where n_g is the number of parents u_1 and v_1 are the hyperparameters and g_i, i = 1,…,6 are the GCA values of 6 parents.

The variance of SCA can be specified as:

$$p(\sigma_s^2 \mid \beta, m_1, m_2, u, W_1, W_2, w_{p1}, w_{p2}, f1, f2, \sigma_e^2, Y)$$
$$\propto (\sigma_s^2)^{-\left(\frac{n_s}{2}+u_2+1\right)} \times \exp\left\{-\frac{1}{\sigma_s^2}\left[\frac{1}{2}\sum_{i=1}^{n_g} s_i^2 + v_2\right]\right\} \propto IG\left[\frac{n_s}{2}+u_2, \frac{1}{2}\sum_{i=1}^{n_g} s_i^2 + v_2\right] \qquad (2.1.6b)$$

Where n_s is the number of crosses and s_i, i=1,……,15 are the SCA values of the 15 crosses.

The error variance component can be specified as:

$$p(\sigma_e^2 \mid \beta, m_1, m_2, u, W_1, W_2, w_{p1}, w_{p2}, f1, f2, \sigma_e^2, Y)$$
$$\propto (\sigma_e^2)^{-\left(\frac{n}{2}+u_3+1\right)} \times \exp\left\{-\frac{1}{\sigma_e^2}\left[\frac{1}{2}(y - X\beta - Zu - W_1 l_1 m_1 - W_2 l_2 m_2)'(y - X\beta - Zu - W_1 l_1 m_1 - W_2 l_2 m_2) + v_3\right]\right\}$$
$$\propto IG\left(\frac{n}{2}+u_3, \frac{1}{2}(y - X\beta - Zu - W_1 l_1 m_1 - W_2 l_2 m_2)'(y - X\beta - Zu - W_1 l_1 m_1 - W_2 l_2 m_2) + v_3\right) \qquad (2.1.6c)$$

2.1.2 Updating of genotypes and genetic parameters: The following updating scheme was used to update the genotypes and the genetic parameters:

1. Initialize θ, σ_g^2, σ_s^2, σ_e^2, a_1, a_2, f_1, f_2 with some reasonable values.
2. Update genotype of parent 1 along with that its offspring with the genotype of other parents known for the first locus using the equations 2.3.2 and 2.3.3. The cycle will repeat for all six parents, keeping the genotypes of other parents constant. The genotype of the progeny is thus updated twice in each cycle, once for each parent.
3. Update the location parameters except for the additive effect of the major gene at the second locus using equation 2.3.5.
4. Update the variances using the equations 2.3.6(a, b and c).
5. Update major gene genotype for parent1 along with its offspring with genotypes of other parents known for the second locus using the equations 2.3.2 and 2.3.3. Repeat the cycle for all the six parents.
6. Update the location parameters except for the additive effect of the major gene at the first locus using equation 2.3.5.
7. Update the variances using the equations 2.3.6(a, b and c).
8. Update the frequencies of favorable alleles at both the loci using equation 2.3.4. Steps 2-8 constitute one iteration in Markov chain Monte Carlo. Updating the parameters twice in each cycle gave more precise estimates of the parameters and improved mixing of the chains. An alternative approach would be to update both the loci simultaneously for each parent in succession.

2.2. Hypotheses to be tested: The model was tested on simulated data sets created with different sets of parameters, with zero, one, or two major genes present and with different values for additive effects of major genes.

The data sets were created using the parameter settings specified in table 2.2.1. The first column specifies the hypothesis which has to be tested using the model, the second column are the parameter settings used to create the simulated data set.

TABLE 2.2.1: **Conceptual hypotheses and the parameter settings to generate simulated data sets (Note: The parameter units are standard phenotypic deviations)**

Conceptual Hypothesis	Parameter settings
1. The model will correctly find no major gene when none is present (purely polygenic model).	Simulate a1=a2=0 $h^2 = 0.2$ $\sigma_s^2/\sigma_a^2 = 0.5$
1a. The model will correctly find no major gene in a purely polygenic model even when the GCA and SCA effects are high.	Simulate a1=a2=0 $h^2 = 0.5$ $\sigma_s^2/\sigma_a^2 = 0.5$
2. The model will correctly find one and only one major gene when one is present.	Simulate a2=0, $h^2 = 0.2$ $\sigma_s^2/\sigma_a^2 = 0.5$ a1 =1.0
3. The model will correctly find two major genes when two are present.	Simulate a1 =1.0, a2=0.5, $h^2 = 0.2$ $\sigma_s^2/\sigma_a^2 = 0.5$
3a. If the additive effects of two major genes are small, the model will correctly find the two major genes.	Simulate a1=0.4 a2 = 0.3

3b. The model will correctly find two major genes when two major genes of equal effect are present.	Simulate a1=a2=1.0

*** If not otherwise stated, f1=f2=0.2, h^2 =0.2, σ_s^2/σ_a^2 = 0.5 .**

2.3. Data generation: The data to evaluate the Extended Mixed Inheritance Model (EMIM) was generated using the 6 parent half-diallel mating design. There were 15 full-sib families, and a total of 2160 progenies in all.

The phenotypic observations of the progeny were simulated using the equation:

$$Y = X\mu + Zu + W_1L_1m_1 + W_2L_2m_2 + e \qquad (2.3.1)$$

Where $X\mu$ is the mean.

Zu = GCA (Parent1) + GCA (Parent2) + SCA.

$W_1L_1m_1$ = Product of multiplication of W1 (matrix of major genotype at locus 1), L_1 (indicator matrix of major gene effects at loci1) and *m1* (major gene effect at first locus).

$W_2L_2m_2$ = Product of multiplication of W_2 (matrix of major genotype at locus 2), L_2 (indicator matrix of major gene effects at loci2) and m_2 (major gene effect at second locus).

e = error term.

The polygenic term was calculated as G1 + G2 + S, where G1 and G2 are the GCA values of the two parents and S is the SCA of the cross of two parents. S has a normal distribution given by N (0, σ_s^2). G_1 and G_2 have a normal distribution given by N (0, σ_g^2). Two parameter heritability (h^2) and ratio of dominance to additive variance (r) were used to calculate the polygenic term. Heritability was defined as $h^2 = 4\sigma_g^2 / \sigma_p^2$ while $r = \sigma_s^2 / \sigma_g^2$. The major genotypes of the parents were assigned as:

$$p(w_p \mid f_1, f_2) = \prod_{i=1}^{n_p} p(w_{pi} \mid f1, f2) \qquad (2.3.2)$$

where w_{pi} is the genotype of the i^{th} parent and *f1* and *f2* are the favorable allele frequencies at loci 1 and 2.

The major genotypes of the progenies were assigned as:

$$p(W \mid w_p) = \prod_{k=1}^{n} p(w_k \mid w_{p1(k)}, w_{p2(k)}) \qquad (2.3.3)$$

where n is the number of progeny per full-sib family.

The expected variance of major genes was calculated as:

$$\sigma_{m1}^2 = 2f_1 (1-f_1) a_1^2 \qquad (2.3.4)$$

$$\sigma_{m2}^2 = 2f_2 (1-f_2) a_2^2$$

The total phenotypic variance is $\sigma_p^2 = \sigma_{m1}^2 + \sigma_{m2}^2 + \sigma_g^2 + \sigma_s^2 + \sigma_e^2 = 1.0$

2.4. Checking the performance of EMIM using the simulated data:

Bayesian output analysis (BOA version 1.5) was used to analyze the results obtained from the MCMC chain. To determine the burn-in-time and convergence Gelman and Rubin shrink factors (Gelman and Rubin, 1992) were used. In addition to these, Brooks, Gelman and Rubin's corrected scale reduction factors and Lewis's dependence factors were used to check for the convergence of the chains. If the dependence factors for a given parameter in a single chain are less than 5 and the corrected scale shrink factors for 0.975 quantiles are less than 1.2, this indicates that chains are converging correctly for those parameters. Dependence factors greater than five indicate convergence failure and we need to reparameterize the model.

Two independent chains for each data set were run to test the consistency of the results from independent runs. The chains were run for 300,000 iterations and the burn-in period for most of the data sets was approximately 30,000 iterations.

The log-likelihood was calculated using the equation 2.4.1. This feature was added when it was found that the two chains were not always consistent with each other. To examine which chain was more reliable mean log-likelihoods were compared. The chain giving more reliable results has a less negative value for the mean log-likelihood compared to the less-reliable chain.

$$likelihood = \frac{1}{2\sigma_e^2}\{(y - X\mu - Zu - W_1 l_1 m_1 - W_2 l_2 m_2)'(y - X\mu - Zu - W_1 l_1 m_1 - W_2 l_2 m_2)\} \quad (2.4.1)$$

CHAPTER III

RESULTS

3.1 No major gene present :

Case (a) No major gene with heritability =0.2: When no major gene was present, the model predicted two major genes with a large effect of around 1.5 SD. The two chains were consistent with each other in the estimated major gene effects and allele frequencies, and gave approximately correct estimates of Ve, Vg and Vs (table 3.1.1).

The two chains gave different parental genotypes for the two loci. The actual genotypes of the parents along with the predicted genotypes for the two chains are given in table 3.1.2. The first chain predicted a heterozygote for parent 5 for both loci, whereas the second chain correctly identified all genotypes as homozygotes for both the loci. The GCA estimates were correspondingly reduced for the cases in which the predicted genotype included favorable alleles. Chain 2 had much higher log likelihood than chain 1.

The estimates for the six GCA values for the two chains are given in table 3.1.3.

TABLE 3.1.1 Means and standard deviations of the five genetic parameters for the two independent chains Run01 and Run02. (Notes: a1 and a2 are the additive effects of major genes at the two loci, Ve,Vg and Vs are the variances of error, GCA and SCA, likelihood is the log likelihood estimates for the two independent chains.**)**

	a1	a2	Ve	Vg	Vs	Likelihood
True	**0**	**0**	**0.898874**	**0.0339067**	**0.019757**	
Run01	-1.5±0.08	1.49+0.08	0.74+0.02	0.086+0.06	0.019+0.04	-767.203
Run02	-1.18+0.16	1.40+0.14	0.73+0.06	0.12+0.15	0.027+0.04	-575.511

TABLE 3.1.2 Genotypes of the parents along with predicted genotype by two independent runs. Genotype 1and Genotype 2 is the genotype of parent at first and second locus respectively.

Parent	Genotype1	Run01	Run02	Genotype2	Run01	Run02
1	**A2A2**	A2A2	A1A1	**B2B2**	B2B2	B2B2
2	**A2A2**	A2A2	A1A1	**B2B2**	B2B2	B2B2
3	**A2A2**	A1A1	A1A1	**B2B2**	B2B2	B1B1
4	**A2A2**	A1A1	A2A2	**B2B2**	B2B2	B1B1
5	**A2A2**	A1A2	A2A2	**B2B2**	B1B2	B2B2
6	**A2A2**	A2A2	A2A2	**B2B2**	B2B2	B2B2

FIGURE 3.1.1 Posterior density of the genotypes at the two loci:
Genotype at the first locus:

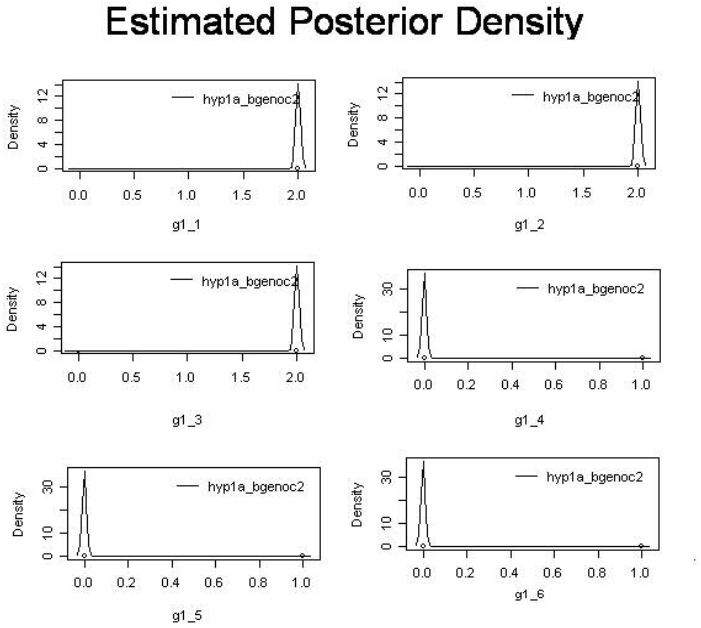


FIGURE 3.1.1 Posterior density of the genotypes at the two loci:

Genotype at the second locus:

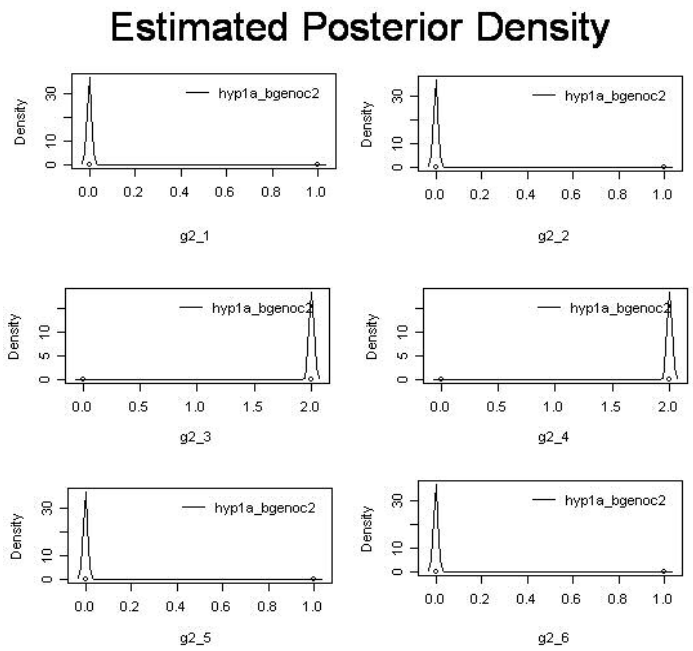


TABLE 3.1.3 GCA of the 6 parents with estimated values by two independent runs Run01 and Run02. The actual values are in bold.

	g1	g2	g3	g4	g5	g6
True	**0.1352**	**0.1217**	**-0.2502**	**-0.0435**	**-0.1335**	**-0.302**
Run01	0.19+0.29	0.18+0.36	-0.22+0.29	-0.02+0.37	-0.13+0.36	-0.21+0.37
Run02	-0.29+0.10	-0.13+0.17	-0.12+0.16	-0.40+0.19	0.09+0.30	-0.16+0.26

TABLE 3.1.4 Raftery and Lewis dependence factors for the five genetic parameters for the two independent runs Run01 and Run02. (Notes: *a1* and *a2* are the additive effects of major genes at the two loci, Ve,Vg and Vs are the variances of error, GCA and SCA **). CSRF are the corrected scale reduction factors for multiple chains based on 0.975 quantiles.**

	a1	a2	Ve	Vg	Vs
Run01	4.903	4.16	1.03	1.01	2.03
Run02	1.88	2.97	2.002	2.04	1.02
CSRF	9.343	1.025	0.422	0.59	1.02

Case (b) No major gene with high GCA values: When no major gene was present the model showed one major gene with additive effect of 0.57, and the variance explained by the major gene increased to 0.16. The second major gene in the model had very small effects. The model assigned equal probability to both the homozygous genotypes except for parent 5, which it incorrectly identified as a heterozygote but only weakly so (p= 0.62; Figure 3.1.2), and hence negatively compensating the GCA estimate.

The estimates of Vg, Vs and Ve were almost identical between the two runs. The estimates of genetic parameters are given in table 1.5.

TABLE 3.1.5. Mean and standard deviations of the five genetic parameters (Notes: a1,a2, f1, f2, Ve, Vg, Vs.a1 and a2 are the additive effect of major gene at two loci, Vg, Vs and Ve are the variances of GCA, SCA and error.) **CSRF are the corrected scale reduction factors for multiple chains based on 0.975 quantiles.**

	a1	a2	Vg	Vs	Ve
True	**0.0**	**0.0**	**0.12**	**0.06**	**0.803**
Run1	0.5753+0.27	0.0967+0.17	0.3457+0.58	0.031+0.03	0.7403+0.081
Run2	0.5728+0.27	0.0954+0.17	0.34+0.60	0.043+0.03	0.74+0.08
CSRF	1.0001	1.0006	1	1.002	1.0003

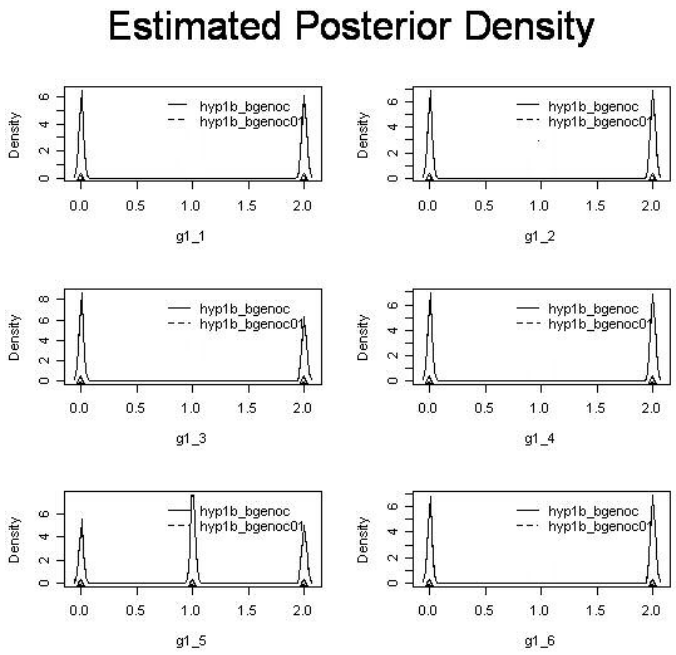
TABLE 3.1.6 Genotypes of the parents along with predicted genotype by two independent runs. Genotype 1and Genotype 2 is the genotype of parent at first and second locus respectively.

Parent	Genotype 1	Run01	Genotype 2	Run01
1	$\mathbf{A_2A_2}$	A_1A_1 (50%), A_2A_2 (50%)	$\mathbf{B_2B_2}$	B_1B_1 (50%), B_2B_2 (50%)
2	$\mathbf{A_2A_2}$	A_1A_1 (50%), A_2A_2 (50%)	$\mathbf{B_2B_2}$	B_1B_1 (50%), B_2B_2 (50%)
3	$\mathbf{A_2A_2}$	A_1A_1 (50%), A_2A_2 (50%)	$\mathbf{B_2B_2}$	B_1B_1 (50%), B_2B_2 (50%)
4	$\mathbf{A_2A_2}$	A_1A_1 (50%), A_2A_2 (50%)	$\mathbf{B_2B_2}$	B_1B_1 (50%), B_2B_2 (50%)
5	$\mathbf{A_2A_2}$	A_1A_2 (60%), A_2A_2 (30%), A_1A_1 (10%)	$\mathbf{B_2B_2}$	B_1B_1 (50%), B_2B_2 (50%)
6	$\mathbf{A_2A_2}$	A_1A_1 (50%), A_2A_2 (50%)	$\mathbf{B_2B_2}$	B_1B_1 (50%), B_2B_2 (50%)

TABLE 3.1.7 GCA estimates of the 6 parents. CSRF are the corrected scale reduction factors for multiple chains based on 0.975 quantiles.

	GCA1	GCA2	GCA3	GCA4	GCA5	GCA6
True	**0.088345**	**-0.45299**	**-0.05679**	**0.1381**	**0.17693**	**0.11972**
Run1	-0.0735±0.63	-0.4169±0.65	-0.0774±0.61	0.0320±0.63	-0.2863±0.63	0.0231±0.63
Run2	-0.0762±0.64	-0.4234±0.65	-0.0693±0.61	0.0257±0.62	-0.2736±0.63	0.0260±0.62
CSRF	1.0007	1.002	0.9999	1.001	1.0005	1.0002

FIGURE 3.1.2 Genotypes of parents for loci with additive effect of 0.5 when no major gene was present:



3.2. One major gene present: The model correctly found one major gene when only one major gene was present. The data was simulated with a1 = 1.0 and a2 = 0.0. The model overestimated the additive effects of major gene at both the loci. The estimate given by the model for a1 and a2 were 1.10 and 0.21. Since the value of a2 was 0.21, one could infer that a second major gene with smaller effect is present. However, all parents are predicted to be homozygous for the second major gene, so a second major gene is not segregating in any of the progeny. The predicted major gene genotypes of the parents at the second locus, which was simulated with additive effect =0.0, are given in figure 3.2.1.

The parameter estimates were precise, with low variances in their posterior distributions (table 3.2.1) The model also gave correct estimates of the genotypes of the parents when the additive effect of the major gene was equal to 1.0 (Figure 3.2.2.,table 3.2.2) The results obtained for two independent chains were consistent with each other.

The Raftery and Lewis dependence factors for single chain were less than 5 (table 3.2.4). Also, the corrected scale reduction factors for multiple chains were less than 1.2, indicating that the chains arose from a stationary distribution. Only the corrected scale reduction factors for the variance of GCA were higher than 1.2. The trace plots of the genetic parameters show that variance of GCA did not vary much in parameter space compared to other genetic parameters. (Figure 3.2.3).

TABLE 3.2.1 Mean and standard deviations of the five genetic parameters (a1,a2, Ve, Vg, Vs). (Notes: a1 and a2 are the additive effect of major gene at two loci, Vg, Vs and Ve are the variances of GCA, SCA and error.) **CSRF are the corrected scale reduction factors for multiple chains based on 0.975 quantiles.**

	a1	a2	Ve	Vg	Vs
True	**1.00**	**0**	**0.6151**	**0.0207**	**0.0.176**
Run01	1.10+0.07	0.21+0.15	0.50+0.22	0.17+0.12	0.07+0.05
Run02	1.22+0.08	0.23+0.10	0.68+0.02	0.09+0.06	0.07+0.05
CSRF	0.872	0.188	1.007	1.57	0.48

TABLE 3.2.2 Genotypes of the parents along with predicted genotype by two independent runs. Genotype 1and Genotype 2 is the genotype of parent at first and second locus respectively.

Parent	Genotype1	Run01	Run02	Genotype2	Run01	Run02
1	**A2A2**	A2A2	A2A2	**B2B2**	B2B2	B2B2
2	**A1A2**	A1A2	A1A2	**B2B2**	B2B2	B2B2
3	**A1A2**	A1A2	A1A2	**B2B2**	B1B1	B1B1
4	**A2A2**	A2A2	A2A2	**B2B2**	B2B2	B2B2
5	**A2A2**	A2A2	A2A2	**B2B2**	B2B2	B2B2
6	**A2A2**	A2A2	A2A2	**B2B2**	B2B2	B2B2

TABLE 3.2.3.GCA estimates of the six parents along with the convergence diagnostic (Notes: Run01 and Run02 are the two independent runs of the same set of data. The actual values are in bold.**) CSRF are the corrected scale reduction factors for multiple chains based on 0.975 quantiles.**

	GCA1	GCA2	GCA3	GCA4	GCA5	GCA6
True	**-0.07483**	**-0.02659**	**0.17816**	**-0.02466**	**-0.16938**	**-0.24066**
Run01	0.076+0.15	0.15+0.20	0.191+0.20	-0.129+0.21	-0.129+0.21	-0.213+0.21
Run02	0.093+0.13	0.099+0.13	0.096+0.17	-0.180+0.09	-0.384+0.16	-0.175+0.20
CSRF	1.04	1.01	0.79	1.05	1.04	1.08

FIGURE 3.2.1 Posterior distribution of the genotypes for locus 2 (estimated additive effect =0.21). The distribution from the run 1 is shown as a solid curve, and that from run 2 is shown as a dashed curve**.**

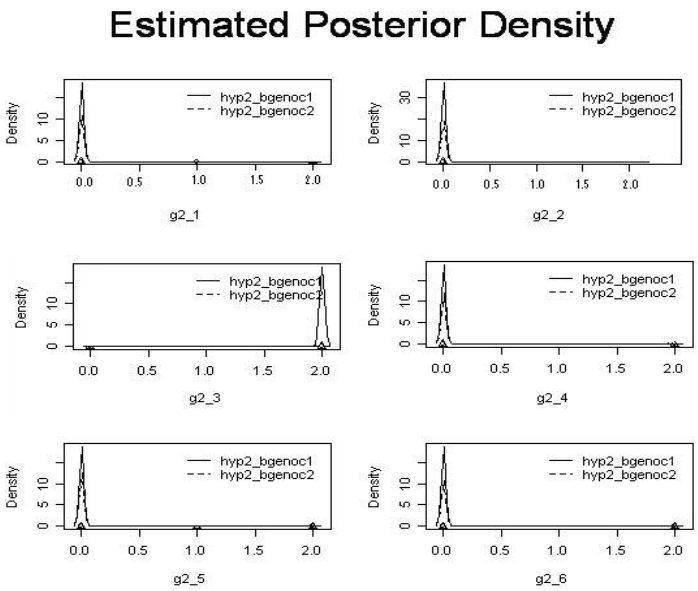


FIGURE 3.2.2. Posterior distribution of the genotypes of the parents for locus 1 (estimated additive effect = 1.10 and 1.22 for the two runs) The distribution from the run 1 is shown as a solid curve, and that from run 2 is shown as a dashed curve**.**

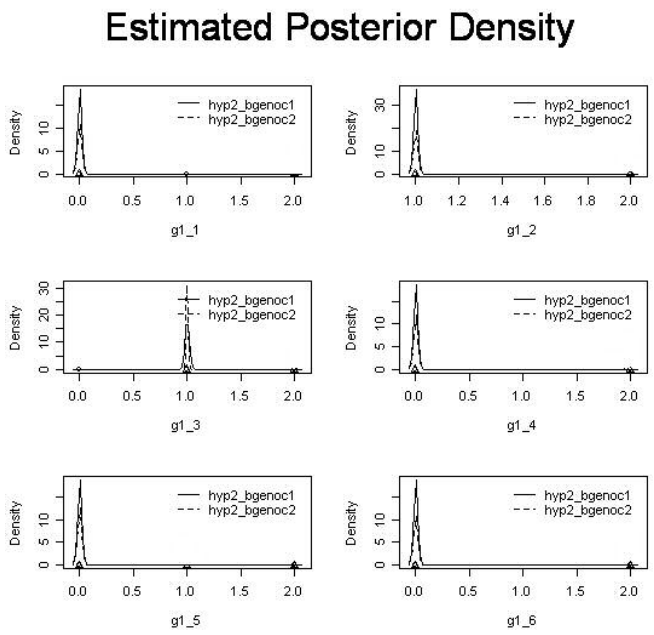
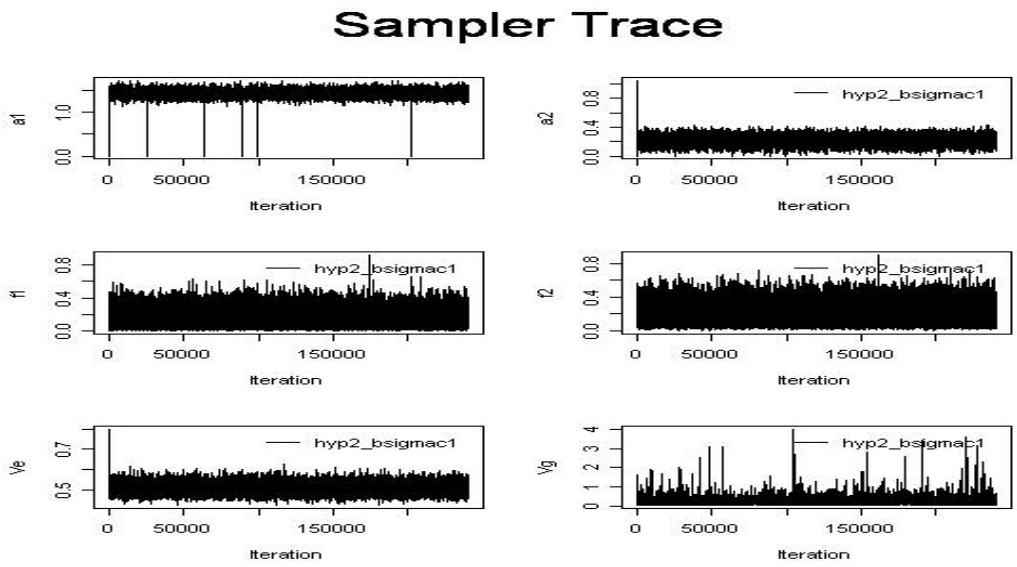


TABLE 3.2.4 Raftery and Lewis dependence factors for the genetic parameters for two independent runs:

Dependence Factors	a1	a2	Ve	Vg	Vs
Run01	4.29	3.22	2.01	1.92	2.02
Run02	3.53	3.79	1.02	1.92	2.07

FIGURE 3.2.3 Trace plots for the 6 genetic parameters (a1, a2, f1, f2, Ve, Vg) for one major gene hypothesis.a1 and a2 are the major gene effect at two loci.



3.3. Two major genes with different effects: The data was simulated with the additive effect of one major gene as 1.0 and the additive effect of the other major gene as 0.5.

The model correctly predicted two major genes; however it overestimated the effects of the two major genes (Table 3.3.1). Thus, the overall variance

explained by the presence of major genes increased from the true value of 0.563 to 1.5 and 1.22 in two different runs (Table 3.3.4).

The model correctly predicted the genotype of the major gene with actual additive effect of 1.0 (Table 3.3.2). For the second major gene with actual additive effect of 0.5, transformations between major gene effects and polygenic effects occurred and the GCA estimates were correspondingly biased. The heterozygotes at the locus with additive effect of 0.5 were not correctly identified and the two runs were not consistent with each other.

TABLE 3.3.1. Mean and standard deviations of the seven genetic parameters (a1,a2, f1, f2, Ve, Vg, Vs). (Notes: a1 and a2 are the additive effect of major gene at two loci, f1 and f2 are the favorable allele frequency, Vg, Vs and Ve are the variances of GCA, SCA and error.) **CSRF are the corrected scale reduction factors for multiple chains based on 0.975 quantiles.**

	a1	a2	f1	f2	Vg	Ve	Vs
True	**1**	**0.5**	**0.33**	**0.41**	**0.026**	**0.501**	**0.033**
Run01	1.30± 0.14	1.5±0.53	0.49±0.18	0.571±0.127	0.149±0.013	0.41±0.36	0.075±0.043
Run02	1.33± 0.14	1.5±0.33	0.35±0.13	0.571±0.123	0.158±0.014	0.49±0.41	0.018±0.049
CSRF	**0.628**	**0.423**	**0.837**	**1**	**1.82**	**1.16**	**1.03**

TABLE 3.3.2. Actual genotypes of the parents along with the genotypes predicted by the model for two independent runs (Notes: Genotype1 is the major gene genotype for the locus with additive effect=1.0, genotype2 is the major gene genotype for the other locus with additive effect =0.5. Run01 and Run02 are the two independent runs for the data.)

	Genotype1	Run01	Run02	Genotype2	Run01	Run02
1	**A2A2**	A2A2	A2A2	**A2A2**	A1A1	A2A2
2	**A2A2**	A2A2	A2A2	**A1A2**	A1A1	A2A2
3	**A1A2**	A1A2	A1A2	**A2A2**	A2A2	A1A1
4	**A1A2**	A1A2	A1A2	**A1A1**	A1A2	A1A2
5	**A1A2**	A1A2	A1A2	**A1A1**	A1A2	A1A2
6	**A1A2**	A1A2	A1A2	**A2A2**	A1A2	A1A2

TABLE 3.3.3 GCA estimates of the six parents along with the convergence diagnostic. CSRF are the corrected scale reduction factors for multiple chains based on 0.975 quantiles.

	g1	g2	g3	g4	g5	g6
Real	**0.0574**	**-0.0014**	**-0.259**	**0.108**	**0.2363**	**0.0756**
Run01	-1.19± 0.18	-1.02±0.18	0.159±0.19	0.276±0.22	0.662±0.18	-0.018±0.20
Run02	-1.37± 0.16	0.309±0.15	0.449±0.17	0.382±0.20	0.651±0.16	-0.156±0.19
CSRF	**0.52**	**1.015**	**0.895**	**1.014**	**0.979**	**2.02**

TABLE 3.3.4. Mean and standard deviations of the variance explained by the major genes at the two loci along with total variance explained by the two loci. Dependence factors are the Raftery and Lewis dependence factors for single chains for the two independent runs 1 and 2:

	Vm	Vm1	Vm2
Actual	**0.56315**	**0.4422**	**0.1209**
Run01	1.56±0.14	0.48±0.04	1.08±0.13
Run02	1.22±0.15	0.82±0.16	0.461±0.16
Dependence Factors			
Run01	1.015	1.012	1.011
Run02	1.001	1.006	0.997

3.4. Two major genes with small additive effects at both loci: The data was simulated with additive effects of the two major genes as 0.4 and 0.3 for the first and second locus, respectively. The model correctly found the two major genes and the estimates of the effect of major genes were accurate for both the loci. However, the model did not correctly predict the genotypes of the parents, and the parents with high GCA were instead shown to have favorable alleles. Consequently the GCA estimates of those parents were biased in a negative direction. The first chain incorrectly predicted homozygote as heterozygote or vice versa in five instances. Mean log-likelihood were used to compare the two chains, and chain 2 had a higher log likelihood (mean log-

likelihood = -117) than the first one (mean log-likelihood = -159). However, chain 2 still had an incorrect assignment of homozygous vs heterozygous genotypes in four instances.

In spite of the difference in the mean log-likelihood and genotypes between the two runs, the trace plots in figure 3.4.1 suggest that the chains mixed well and the parameter values moved well in sample space except for the variance of GCA. The dependence factors for single chains were less than 5.0 except for the variance of GCA, for which the dependence factors were greater.

TABLE 3.4.1. Mean and standard deviations of the five genetic parameters (*a1*, a2, Ve, Vg, Vs, likelihood is the log likelihood estimates for the two independent chains.)

	a1	a2	Ve	Vg	Vs	Likelihood
True	**0.4**	**0.3**	**0.57**	**0.0321**	**0.02256**	
Run01	0.405+0.09	0.306+0.10	0.426+0.22	0.26+0.17	0.079+0.07	-159.57
Run02	0.42+0.04	0.29+0.08	0.39+0.25	0.33+0.23	0.066+0.06	-117.75
RDF	4.05	3.28	1.92	9.72	1.96	

TABLE 3.4.2. Genotypes of the parents, the actual genotypes at the two loci are indicated with bold letters. Run01 and Run02 are the estimates of parental genotypes given by the model:

	Genotype1	Run01	Run02	Genotype 2	Run01	Run02
1	**A1A2**	A1A1	A1A2	**A1A2**	A1A1	A1A2
2	**A2A2**	A1A2	A1A1	**A1A2**	A1A2	A2A2
3	**A1A2**	A1A2	A2A2	**A2A2**	A1A2	A1A1
4	**A2A2**	A2A2	A2A2	**A2A2**	A2A2	A1A1
5	**A2A2**	A2A2	A1A2	**A1A2**	A2A2	A1A2
6	**A1A2**	A1A2	A2A2	**A1A2**	A1A2	A1A2

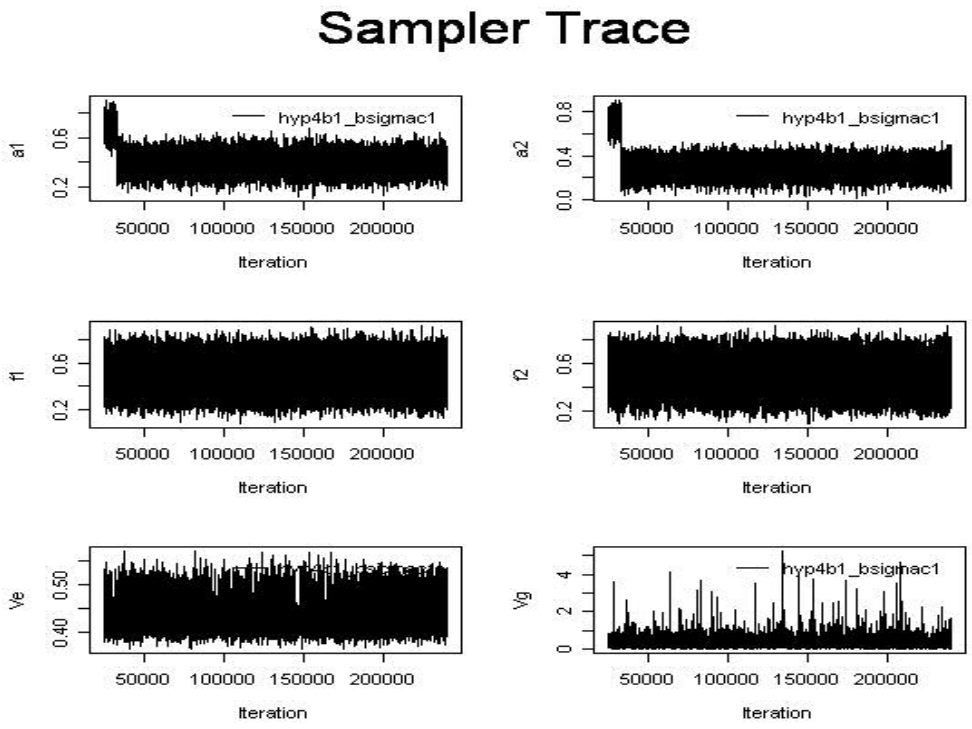
TABLE 3.4.3 GCA estimates of the six parents along with the convergence diagnostic. **CSRF are the corrected scale reduction factors for multiple chains based on 0.975 quantiles.**

	GCA1	GCA2	GCA3	GCA4	GCA5	GCA6
True	***0.18536***	**0.16423**	**0.3052**	**-0.15955**	**0.0087**	**-0.0933**
Run01	-0.6191±0.17	-0.1992+0.16	0.2454+0.18	-0.7257+0.24	0.7586+0.28	0.1727+0.21
Run02	-0.201+0.19	0.786+0.17	-0.311+0.21	-0.245+0.25	-0.064+0.19	-0.40+0.23
CSRF	0.821	1.245	1.231	0.45	2.08	1.92

TABLE 3.4.4 Mean and standard deviations of the variance explained by the major genes at the two loci along with total variance explained by the two loci(Notes: There were two separate runs Run01 and Run02. RDF is the Raftery and Lewis dependence factors for single chain, in this case Run01 RDF are shown.)

	Vm	Vm1	Vm2
True	**0.099**	**0.06**	**0.039**
Run01	0.129±0.06	0.080±0.04	0.048±0.04
Run02	0.121±0.05	0.079±0.03	0.042±0.05
RDF	0.98	3.907	3.209

FIGURE 3.4.1.Trace plots for the six genetic parameters (a1, a2, f1, f2, Ve, Vg)



3.5. Two major genes with additive effect of major genes large and equal:

The main concern in this case was to find out whether the model would correctly distinguish the effects of two major genes when their effects are equal. The data in this case was simulated with both a1=a2=1.0. The model correctly found two major genes with equal effects, and correctly predicted the genotypes of the parents at both the loci (table 3.5.2). The model gave accurate estimates for the genetic parameters and the two independent runs were consistent with each other (table 3.5.1).

The chains converged well; the Raftery and Lewis dependence factors for single chains were less than 5.0 as well as the corrected scale reduction factors (Gelman and Rubin shrink factors) for multiple chains were less than 1.2 (table 3.5.3).

TABLE 3.5.1. Mean and standard deviations of the seven genetic parameters (a1,a2, f1, f2, Ve, Vg, Vs)

	a1	a2	f1	f2	Ve	Vg	Vs
True	**1**	**1**	**0.167**	**0.167**	**0.281562**	**0.064515**	**0.018299**
Run01	1.19+0.07	1.17+0.07	0.14+0.09	0.15+0.09	0.42+0.04	0.31+0.24	0.07+0.06
Run02	0.90+0.13	1.00+0.05	0.12+0.12	0.15+0.09	1.00+0.77	0.22+0.12	0.067+0.044

TABLE 3.5.2. Actual and predicted genotypes of the parents at the two loci (genotype1 is the genotype of major gene at first loci, genotype2 is the genotype at loci 2)

	Genotype1	Run01	Run02	Genotype2	Run01	Run02
1	$\mathbf{A_1A_2}$	A_1A_2	A_1A_2	$\mathbf{B_2B_2}$	B_2B_2	B_2B_2
2	$\mathbf{A_2A_2}$	A_2A_2	A_2A_2	$\mathbf{B_2B_2}$	B_2B_2	B_2B_2
3	$\mathbf{A_2A_2}$	A_2A_2	A_2A_2	$\mathbf{B_1B_2}$	B_1B_2	B_1B_2
4	$\mathbf{A_1A_2}$	A_1A_2	A_1A_2	$\mathbf{B_2B_2}$	B_2B_2	B_2B_2
5	$\mathbf{A_2A_2}$	A_2A_2	A_2A_2	$\mathbf{B_2B_2}$	B_2B_2	B_2B_2
6	$\mathbf{A_2A_2}$	A_2A_2	A_2A_2	$\mathbf{B_1B_2}$	B_1B_2	B_1B_2

FIGURE 3.5.1.Posterior densities of the genotypes of the parents for the two loci

Genotype of the parents at first loci:

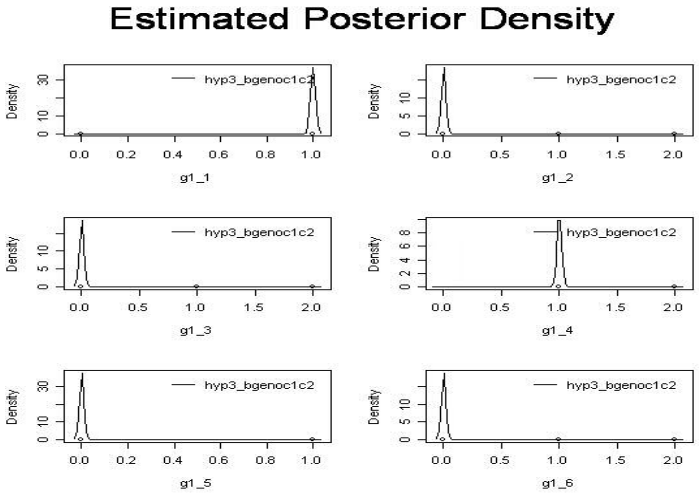


FIGURE 3.5.1.Posterior densities of the genotypes of the parents for the two loci
Genotype of the parents at second loci:

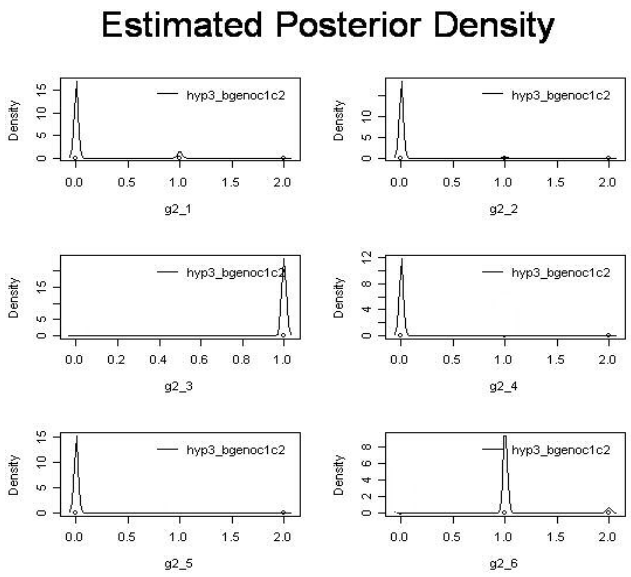


TABLE 3.5.3 Raftery and Lewis dependence factors for the genetic parameters for two independent runs. CSRF is the corrected scale reduction factors for multiple chains based on 0.975 quantiles.):

	a1	a2	Ve	Vg	Vs
Run01	2.341	3.205	1.099	2.639	4.12
Run02	1.889	2.659	1.381	1.927	2.091
CSRF	0.319	0.482	0.992	0.129	0.383

CHAPTER IV

DISCUSSION

Gibbs sampling with parent blocking was used to obtain the posterior distribution of genetic parameters along with major gene genotypes at the two loci. The model correctly predicted major gene genotypes and the genetic parameters when major gene effects were large (additive effect ≥ 1.0 phenotypic standard deviation units). However, when major gene effects were small (additive effect $\leq$ 0.5), transformation between polygenic effects and major gene effects occurred frequently. More importantly, homozygous parents were often mistakenly identified as heterozygotes and vice versa when major gene effects were $\leq$ 0.5. This issue is not specific to the two-gene model because similar results were obtained from the Zeng et al (2003) single gene model in which the model incorrectly assigned the genotypes of the parents when the effects of major gene were $\leq$ 0.5.

For each hypothesis, two independent chains were run, each with 300,000 iterations. The burn-in period required for achieving a stationary distribution was around 30,000 iterations for most of the data sets. However, when the additive

effects of major genes were large and equal (a1=a2=1.0) the required burn-in period was around 90,000 iterations.

The two chains were compared with each other and with the actual simulated data sets to evaluate the performance of the model. In most cases the two chains were consistent with each other. In cases where the two chains were not consistent with each other, mean log likelihood estimates were used to compare the two chains, and the chain with the less negative log likelihood was considered to be more reliable. If the difference between the mean log likelihoods of independent chains was large, this gave another indication that the two chains were not converging to the same values and were not mixing well. The diagnostics alone did not provide very reliable results in this case, and the parental genotypes and mean log-likelihood estimates provided a better indication of whether the chains were mixing properly and converging to the same distributions. When the major gene effects were large (*a1*=*a2*=1.0), the model assigned correct genotypes to the parents and estimates of genetic parameters were precise and reasonably accurate. The two chains gave consistent results with each other for all the genetic parameters.

When the additive effects of major genes were small (0.4 and 0.3) the model correctly estimated the effects of major genes, but the two chains were neither consistent nor correct in the assignment of the genotypes. Parent with high GCA values were instead inferred to have favorable alleles at the major genes in several instances. The mean log likelihoods of the two chains were

substantially different, but even the chain with higher log likelihood had incorrect homozygous vs heterozygous genotype assignments.

When only one major gene was present the model identified two major genes one with the effect of 1.0 and the other with the effect of 0.21. However, all predicted genotypes for second locus were homozygotes, thus correctly indicating that a second major gene was not segregating.

The model incorrectly predicted two major genes with large effects when no major genes were present. One of the chains identified one parent as a heterozygote for both loci. The second chain which had much higher log likelihood correctly identified all genotypes as homozygotes. In other respects, the two chains were consistent with each other and correctly estimated the variance of error, GCA and SCA.

When no major gene was present but the data was simulated with high heritability of polygenic effects (h^2 = 0.5), the model found one major gene with additive effect of 0.57. Since the model assigned equal probability to both homozygous genotypes in all but one parent, where the posterior probability was not very high for the heterozygote (0.62), the results do not provide strong support of major gene effects. Similar results were obtained from one gene model by Zeng et al (2003). However in their case the model pulled up all homozygotes, thus indicating that no major gene was present.

One of the main issues with all MCMC procedures is whether the chains are converging or not. From the corrected scale reduction factors and the trace

plots obtained for the parameters, we could conclude that the chains mixed well and the parameter estimates came from a stationary distribution. The corrected scale reduction factors (CSRF) for variance of GCA were higher than 1.2 in some cases, but otherwise all the parameters appeared to arise from a stationary distribution with CSRF much less than 1.2. Although the diagnostic suggested that the parameters arose from stationary distribution, the difference in mean log-likelihood of chains and the parental genotypes indicate that the two chains did not mix well in some of the data sets. Thus, the mean log-likelihoods were a more useful tool to analyze the results as compared to the convergence diagnostic provided by BOA.

In application of this model, multiple chains for the same set of data should be run (e.g. 3 to 5 chains rather than the two replicates we used) and the results obtained should be compared to each other along with the likelihood estimates. If chains converge to the same parameter estimates, including parental genotypes, and the likelihood estimates of the chains are not different then the results are more likely to be dependable. If the chains converge on different parental genotypes and the mean log-likelihoods of the chains are different, our results suggest the chains with higher log likelihoods are at least closer to identifying the correct genotypes.

The ability to identify parents that are heterozygous for major genes may be useful in plant breeding programs, in combination with more conventional quantitative genetic criteria such as GCA estimates. In a more general sense,

the model developed here could potentially be a valuable aid in identifying parents that are heterozygous for major genes in studies of adaptive genetic variation in natural populations. With only the phenotypic observations of progeny, the genotype of the parents and the size of major gene effects could be estimated. When two major genes were actually present, the model performed better than the single gene model by Zeng et al (2003), which underestimated the size of major gene effects when two major genes were present and the values of *a* and *Vm* were half of the actual values. The cost of mapping without knowledge of whether a particular cross is segregating for major QTL could thus be avoided using this technique, and mapping efforts could be directed toward the crosses most likely to be segregating for major genes. However, we found the model to be reliable only when the major gene effects were large ($2a > 1.0$ standard deviations). With careful examination of the mean log-likelihood estimates of the chains and multiple runs of the same data set one may be able to detect the genes with smaller effect. It has been shown that when the major gene effects were low (case 3.4), the two independent runs gave consistent and accurate estimates for the major gene effects, but the heterozygotes were incorrectly identified.

When no major gene or one major gene was segregating, the model performed comparably to Zeng's (2003) single gene model. Both the models predicted major gene effects when no major genes were segregating, but did not show a strong posterior probability for the presence of heterozygotes.

In summary, the two models, i.e. the single gene model by Zeng (2003) and the two gene model, performed equally well when one major gene was present. When no major gene was present and the heritability was high, both the models predicted one major gene with effect of 0.5 S.D. However the single gene model predicted all homozygotes, whereas the two gene model identified one parent as a heterozygote with weak posterior probability (60%). Thus, the two gene model still does not provide strong evidence for segregation of a major gene. When two major genes were present the single gene model underestimated the effects of major genes, whereas the two gene model correctly estimated the effects of the two genes in all but one case, in which it overestimated the effects of one major gene.

Both the Zeng (2003) single gene model and the two gene model had problems with mixing. To improve the performance of the model, it may be useful to explore a more complex algorithm like Metropolis-Hastings to get the posterior distribution of the parameters and the genotypes. Another approach would be to update progeny genotypes one allele at a time when the genotype of each individual parent is updated, while treating the allele from the other parent as known, rather than updating the entire genotype of each progeny twice for each cycle (equation 3.2). Under this approach, the mixture likelihood model of progeny genotypes from which the parent genotypes are sampled would also treat the allele from the other parent as known (equation 3.1). This approach has

not been tested in the present study, so its effect on the mixing of chains is unknown.

In this model only the additive effects of major genes were taken into account; a further modification could be addition of dominance effects to the model. Also, more complex interactions, such as genotype X environment interactions and epistatic interactions could be included in future models. Finally, one could also model the effects of linkage disequilibrium between the genes.

REFERENCES

Beavis, W. D., 1994 The power and deceit of QTL experiments: lessons from comparative QTL studies, pp. 250–266 in Proceedings of the 49th Annual Corn and Sorghum Industry Research Conference. American Seed Trade Association, Washington, DC.

Bogdan M., J. K. Ghosh, and R. W. Doerge Modifying the Schwarz Bayesian Information Criterion to Locate Multiple Interacting quantitative Trait Loci Genetics, June 1, 2004; 167(2): 989 - 999.

Carson, M.L., Stuber, C.W., Senior, M.L. 2004. Identification and mapping of quantitative trait loci conditioning resistance to southern leaf blight of maize caused by Cochliobolus heterostrophus race O. Phytopathology. 94:862-867.

Dempster, A. P., N. M. Laird, and D. B. Rubin, 1977 Maximum likelihood from incomplete data via the EM algorithm. Journal R. Stat. Soc. 39:1-38.

Dupuis, J. and D. Siegmund, 1999 Statistical methods for mapping quantitative trait loci from a dense set of markers. Genetics 151:373-386.

Feenstra B.and I. M. Skovgaard A Quantitative Trait Locus Mixture Model That Avoids Spurious LOD Score Peaks Genetics, June 1, 2004; 167(2): 959 - 965.

Fry, J.D., K.A. deRonde and T.F.C. Mackay. 1995. Polygenic mutation in Drosophila melanogaster: genetic analysis of selection lines. Genetics 139:1293-1307.

Grignola, F. E., I. Hoeschelle, and B. TIER, 1996a Mapping quantitative trait loci via residual maximum likelihood: II. A simulation study. Genet. Sel. Evol. 28:479-490.

Hackett, C. A. and J. I. Weller, 1995 Genetic mapping of quantitative trait loci for traits with ordinal distributions. Biometrics 51:1252-1263.

Haley, C. S., S. A. Knott, and J. M. Elsen, 1994 Mapping quantitative trait loci in crosses between outbred lines using least squares. Genetics 136:1195-1207.

Haley, C. S. and S. A. Knott, 1992 A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity 69:315-324

Hoeschelle, I. and P. Vanranden, 1993a Bayesian analysis of linkage between genetic markers and quantitative trait loci: I. Prior knowledge. Theor. Appl. Genet. 85:953-960.

Hoeschelle, I. and P. Vanranden, 1993b Bayesian analysis of linkage between genetic markers and quantitative trait loci: II. Combining prior knowledge with experimental evidence. Theor. Appl. Genet. 85:946-952.

Jiang, C.J. et al. (1994). The use of mixture models to detect effects of major genes on
quantitative characters in a plant breeding experiment. Genetics, 136, 383-394.

Kao, C.H., 1995 Statistical methods for locating the positions and analyzing epistasis of multiple quantitative trait genes using molecular marker information. Ph.D. Thesis, North Carolina State University, Raleigh.

Kao C.H.On the Differences Between Maximum Likelihood and Regression Interval Mapping in the Analysis of Quantitative Trait Loci
Genetics, October 1, 2000; 156(2): 855 - 865.

Kao C.H. and Z.B. Zeng Modeling Epistasis of Quantitative Trait Loci Using Cockerham's Model Genetics, March 1, 2002; 160(3): 1243 - 1261.

Kaya, Z., Sewell, M.M. and Neale, D.B. (1999). Identification of quantitative trait loci
influencing annual height- and diameter increment growth in loblolly pine (Pinus taeda L.). Theor. Appl. Genet., 98,586-592.

Lander, E. and L. Kruglyak, 1995 Genetic dissection of complex traits: guidelines for intrepreting and reporting linkage results. Nat. Genet. 11:241-247

Levin I, Cahaner A, Rabinowitch H.D., Elkind Y Effects of the MS10 gene, polygenes and their interaction on pistil and anther-cone lengths in tomato flowers Heredity 73: 72-77 Part 1, Jul 1994.

Louis, T. A., 1982 Finding the observed information matrix when using the EM algorithm. J. R. Stat. Soc. Ser. B 44:226-233.

Mather K.Polygenic inheritance and natural selection Biological Reviews 18:32-64

Otto S. P.and C. D. Jones Detecting the Undetected: Estimating the Total Number of Loci Underlying a Quantitative Trait Genetics, December 1, 2000; 156(4): 2093 - 2107.

Piper, L.R. and Shrimpton, A.E. (1989). The quantitative effects of genes which influence metrics traits. In: Hill, W.G., Mackay, T.F.C. (eds) Evolution and animal breeding. Reviews on molecular and quantitative approaches in 125 honor of Alan Robertson. Wallingford, CBA Int, 147-151.

Satagopan, J. M., B. S. Yandell, M. A. Newton, And T. C. Osborn, 1996 A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. Genetics 144:805-816

Sen and G. A. Churchill A Statistical Framework for Quantitative Trait Mapping Genetics, September 1, 2001; 159(1): 371 - 387.

Sewell, M. M., D. L. Bassoni, R. A. Megraw, N. C. Wheeler, and D. B. Neale, 2000 Identification of QTLs influencing wood property traits in loblolly pine (Pinus taeda L.). I. Physical wood properties. Theor. Appl. Genet. 101:1273-1281.

Sewell, M. M., M. F. Davis, G. A. Tuskan, N. C. Wheeler, and C. C. Elam et al., 2002 Identification of QTLs influencing wood property traits in loblolly pine (Pinus taeda L.). II. Chemical wood properties. Theor. Appl. Genet. 104:214-222.[Medline]

Sillanpaa, M. J. and E. Arjas, 1998 Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. Genetics 148:1373-1388[Abstract/Free Full Text].

Simko, Ivan, S Costanzo, KG Haynes, BJ Christ and RW Jones. "Linkage disequilibrium mapping of a Verticillium dahliae resistance QTL in tetraploid potato (Solanum tuberosum) through a candidate gene approach.Am J. Of Potato Research Vol. 81, No. 1, Pgs. 88-89

Strauss, S. H., R. Lande, and G. Namkoong, 1992 Limitations of molecular-marker-aided selection in forest tree breeding. Can. J. For. Sci. 22:1050-1061.

Tanksley, S.D. (1993). Mapping genes. Annual Review of Genetics 27, 205-233.
Terasvirta, T. and I. Mellin, 1986 Model selection criteria and model selection tests in regression models. Scand. J. Stat. 13:159-171.

Thaller, G. and I. Hoeschele, 1996 A Monte Carlo method for Bayesian analysis of linkage between single markers and quantitative trait loci: I. Methodology. Theor. Appl. Genet. 93:1161-1166.

Uimari, P. and I. Hoeschele, 1997 Mapping linked quantitative trait loci using Bayesian method analysis and Markov chain Monte Carlo Algorithms. Genetics 146:735-743.

Uimari, P., G. Thaller, and I. Hoeschele, 1996 The use of multiple markers in a Bayesian method for mapping quantitative trait loci. Genetics 143:1831-1842.

Ungerer, M. C., S. S. Halldorsdottir, J. L. Modliszewski, T. F. C. Mackay, and M. D. Purugganan. 2002. Quantitative trait loci for inflorescence development in *Arabidopsis thaliana*. Genetics 160:1133-1151.

Ungerer, M. C., S. S. Halldorsdottir, M. D. Purugganan, and T. F. C. Mackay. 2003. Genotype-environment interactions at quantitative trait loci affecting inflorescence development in *Arabidopsis thaliana*. Genetics 165:353-365.]

Weber K., Robert Eisman,Shawn Higgins, Lisa Morey,April Patty,Michele Tausek and Zhao-Bang Zeng An Analysis of Polygenes Affecting Wing Shape on Chromosome 2 in Drosophila melanogaster Annu. Rev. Genet. 27:205–233.

Wright, F. A. and A. Kong, 1997 Linkage mapping in experimental crosses: the robustness of single-gene models. Genetics 146:417-425.

Xu, S., 1995 A comment on the simple regression method for interval mapping. Genetics 141:1657-1659[Free Full Text].

Yi N., S. Xu, and D. B. Allison Bayesian Model Choice and Search Strategies for Mapping Interacting Quantitative Trait Loci Genetics, October 1, 2003; 165(2): 867 - 883.

Yi N. and S. Xu Bayesian Mapping of Quantitative Trait Loci Under Complicated Mating Designs Genetics, April 1, 2001; 157(4): 1759 - 1771.

Yi N. and S. Xu Bayesian Mapping of Quantitative Trait Loci for Complex Binary Traits Genetics, July 1, 2000; 155(3): 1391 - 1403.

Zeng, Z.B., 1994 Precision mapping of quantitative trait loci. Genetics 136:1457-1468.

Zeng W. Statistical methods for detecting major genes of quantitative traits using phenotypic data of diallel mating. PhD Dissertation North Carolina State University, Raleigh.(2000)

Zeng W., Ghosh S., Li B. A blocking Gibbs sampling method to detect major genes with phenotypic data from diallel mating. Genet. Res. Camb.(2004) 143-154.

Zou F., B. S. Yandell, and J. P. Fine Rank-Based Statistical Methodologies for Quantitative Trait Locus Mapping Genetics, November 1, 2003; 165(3): 1599 - 1605.